Exploiting the Internet to build language resources for less resourced languages

Antton Gurrutxaga, Igor Leturia, Eli Pociello, Iñaki San Vicente, Xabier Saralegi
Elhuyar Foundation
Zelai Haundi kalea 3
Osinalde Industrialdea
20170 Usurbil, Spain
E-mail: {a.gurrutxaga, i.leturia, e.pociello, i.sanvicente, x.saralegi }@elhuyar.com

Abstract

This paper aims to present a general view of the Elhuyar Foundation's strategy to build several types of language resources for Basque out of the web in a cost-efficient way. We collect various types of corpora (specialized, general, comparable, parallel…) from the Internet by means of automatic tools, and then other kinds of resources (terminology, ontologies, etc.) are built out of them also using other automatic tools that we have developed. We have also built a web-as-corpus tool to query the web directly as if it were a corpus. In the end of the paper, we describe two experiments that we have performed to prove the validity of the approach: one that automatically collects specialized corpora in Basque and extracts terminology out of them, and another one that automatically collects a comparable corpus and extracts bilingual terminology out of it, using web-derived contexts to improve the results. In our opinion, the strategy is very interesting and attractive for other less resourced languages too, provided they have enough presence on the web.

1. Motivation

Any language aiming to survive in a world that is becoming more intercommunicated and global day by day, and to be used normally in education, media, etc., must necessarily have at its disposal language resources such as dictionaries or corpora, preferably in digital form. The ever-growing presence of ICTs in everyday life adds to these requisites the existence of language technologies and NLP tools for that language, which in turn also need electronic dictionaries and corpora in order to be developed. Therefore, the need for lexical resources and corpora of any language intending to be modern is undeniable.

Besides, modern lexicography and terminology is hardly done based solely on experts' knowledge or intuition; empirical evidence is needed or previous use at least is studied, and these are provided by corpora. And there are many tools that ease the process of building lexical or terminological dictionaries by making use of NLP and statistical methods to automatically extract candidates out of corpora.

So it is clear that corpora of any kind (monolingual, parallel, comparable...) are a very valuable resource for many aspects of the development of a language. And generally, the bigger the corpora, the better the results obtained from them. But less resourced languages are not exactly rich in corpora, let alone big corpora: on the one hand, building a corpus in the classical way, i.e. out of printed texts, is normally a very costly process; on the other, the number of language experts or researchers dealing with these languages is much smaller than that of major languages.

However, the Internet provides a huge number of texts in a digital and easy to manipulate standard format. For any less resourced language there are bound to be many more texts on the web than in any corpus. That is why turning to the Internet to build corpora (and, through them, other kinds of resources such as dictionaries, terminology lists or statistical machine translation systems) is a very attractive and logical choice for less resourced languages.

The Elhuyar Foundation has been exploring this path for the last few years in order to build language resources for the Basque language. In the following sections we will explain the problems we have encountered and the approaches we have followed for each kind of resource, the former presumably being similar to those that other less resourced languages might encounter, and the latter hopefully being applicable to them too.

2. Using the web to build corpora

2.1 Monolingual specialized corpora

Specialized corpora, that is, corpora made out of texts belonging to a certain domain or topic, are a very valuable resource for terminology tasks as well as for most NLP tasks. Major languages often build specialized corpora by simply crawling one website, or a few, dedicated to the topic and which contain a large number of texts on it. Sometimes this method is combined with some machine-learning filter tailor-made for the specific topic, in order to follow links to external sites, too. But for Basque (and most likely for many other less-resourced languages) there are not many websites that are specialized in a topic and which contain a significant number of texts, or at least there are not for any topic one can think of. And the process of building machine-learning filters is too costly due to the lack of training data.

Hence, for Basque a whole web-wide approach must be used, using search engines. The *de facto* standard process major languages use for collecting web-wide specialized corpora, which was first used by the BootCaT tool (Baroni & Bernardini, 2004), consists of starting from a given list of words, asking APIs of search engines for random combinations of them and downloading the returned pages. However, the topic precision that can be obtained by this methodology has scarcely been measured, and a small evaluation performed on the original BootCaT paper hints that one third of the texts could be unrelated to the topic. And this precision is much worse when searching for corpora in the Basque language. Some experiments we have performed show that this can drop to only 25% (Leturia et al., 2008a).

The main reasons for this are two: one is that no search engine offers the possibility of returning pages in Basque alone, so when looking for technical words (as is often the case with specialized corpora), it is very probable that they exist in other languages too, and thus the queries return many pages that are not in Basque; the other is that Basque is a morphologically rich language and any lemma has many different word forms, so looking for the base form of a word alone, as search engines do, brings fewer results.

Many other languages suffer from these problems regarding search engines. Less than fifty languages are treated properly by Google, Yahoo or Bing. In the case of Basque, we have solved them to some extent (Leturia et al., 2008b). For the former, we use the language-filtering words method, consisting of adding the four most frequent Basque words to the queries within an AND operator, which raises language precision from 15% to over 90%. For the latter, we solve it by means of morphological query expansion, which consists of querying for different word forms of the lemma, obtained by morphological generation, within an OR operator. In order to maximize recall, the most frequent word forms are used, and recall is improved by up to 60% in some cases.

These two techniques raise the topic precision to the baseline of other languages (roughly 66%). Nevertheless, we have developed a method to try to further improve topic precision and have implemented it in a system to automatically collect Basque specialized corpora from the Internet called AutoCorpEx (Leturia et al., 2008a). Its operation is explained below.

The system is fed with a sample mini-corpus of documents that covers as many sub-areas of the domain as possible –10-20 small documents can be enough, depending on the domain. A list of seed terms is automatically extracted from it, which can be manually edited and improved if necessary. Then combinations of these seed words are sent to a search engine, using morphological query expansion and language-filtering words to obtain better results for Basque, and the pages returned are downloaded. Next, the various cleaning and filtering stages necessary in any corpus collecting process involving the web are performed. Boilerplate is stripped off the downloaded pages (Saralegi and Leturia, 2007) which are then passed through various filters: size filtering (Fletcher, 2004), paragraph-level language filtering, near-duplicate filtering (Broder, 2000) and containment filtering (Broder, 1997). After that we have added a final topic-filtering stage, using the initial sample mini-corpus as a reference and employing document similarity techniques (Saralegi and Alegria, 2007) based on keyword frequencies (Sebastiani, 2002). A manual evaluation of this tool showed that it can obtain a topic precision of over 90%.

2.2 Multilingual domain-comparable corpora

Multilingual corpora are considered comparable if the subcorpora of each of the different languages share some common feature, such as domain, genre, time period, etc. Specifically, the texts of a domain-comparable corpora are all in the same domain. These kinds of resources are very useful for automatic terminology extraction, statistical machine translation training, etc., although they are more difficult to exploit than parallel corpora (because of their smaller alignment level, there is less explicit knowledge to extract). However, parallel corpora of significant size are scarce, especially for less resourced languages, and since comparable corpora are easier to obtain, more and more research is heading towards the exploitation of these kinds of corpora.

With the method described in section 2.1 for collecting monolingual specialized corpora, domain-comparable corpora can also be built (Leturia et al., 2009): we can use a sample mini-corpus for each language and launch the corpus collecting process independently for each of them; if the sample mini-corpora that are used for the domain filtering are comparable or similar enough (ideally, a parallel corpus would be best), the corpora obtained will be comparable to some extent, too. We have implemented this methodology in a tool called Co3 (Comparable Corpora Collector).

We have also developed and tried another variant of this method; it uses only a sample mini-corpus in one of the languages, and translates the extracted seed words (they are manually revised) and the keyword vectors used in the domain-filtering to the other language by means of a bilingual dictionary.

This method, theoretically, presents two clear advantages: firstly, the sample mini-corpora are as similar as can be (there is only one), so we can expect a greater comparability in the end; and secondly, we only need to collect one sample corpus. However, it presents some problems too, mainly the following two: firstly, because dictionaries do not cover all existing terminology, we may have some Out Of Vocabulary (OOV) words and the method may not work so well; secondly, we have to deal with the ambiguity derived from dictionaries, and selecting the right translation of a word is not so easy. To reduce the amount of OOV words, the ones that have been POS-tagged as proper nouns are included as they are in the translated lists, since most of them are named entities. And for resolving ambiguity, for the moment, we have used a naïve “first translation” approach, widely used as a baseline in NLP tasks that involve translation based on dictionaries. An evaluation showed that the results of the dictionary-based method were no worse than those of the two sample mini-corpora method.

2.3 Monolingual general corpora

The web is also used as a source for large general corpora, which are very interesting for tasks such as language standardization, general lexicography, discourse analysis, etc. Again, two approaches exist, one based on crawling and the other on search engines. The crawling method is used in the projects of the WaCky initiative (Baroni et al., 2009), which have collected gigaword-size corpora for German (Baroni and Kilgarriff, 2006), Italian (Baroni and Ueyama, 2006) and English (Ferraresi et al., 2008), with many others on the way. Search engines are used for example by Sharoff (2006), sending combinations of the 500 most frequent words of the language.

Currently, we have ongoing projects for collecting large general corpora for Basque using both methods. The usual cleaning and filtering is done in all cases, and the search engine-based approach uses the aforementioned morphological query expansion and language-filtering words techniques. So far, the crawling-based method has gathered a 250-million-word corpus and the search engine-based method a 100 million word corpus.

2.4 Other kinds of corpora

We have already mentioned that parallel corpora (multilingual corpora made out of texts that are translations, preferably aligned at the sentence level, such as translation memories) are very useful for machine translation, terminology extraction, etc., but are not easy to obtain. However, the web is full of websites with versions in more than one language; specifically, most corporate or public websites that are in a less resourced language also include a version in one or more major languages. This fact has already been exploited for automatically building parallel corpora (Resnik, 1998). In the same line of work, we have an ongoing project, called PaCo2 (Parallel Corpora Compiler) to automatically collect Basque-Spanish or Basque-English parallel corpora from the Internet.

For the near future, we also have an interest in genre-specific corpora. *A priori*, we can expect to be able to collect these kinds of corpora by crawling, at least for some genres such as journalism, blogs, administration, since there are websites with large amounts of content of those genres. For others, genre filters or classifiers would have to be developed. Such tools have been built for major languages, which use punctuation signs or POS trigrams as filtering features (Sharoff, 2006); tests have yet to be carried out to see whether these features work for an agglutinative language like Basque.

3. Building other kinds of resources

3.1 A web-as-corpus tool

A common use of corpora is to use them for linguistic research: querying for one or more words and looking at their counts, contexts, most frequent surrounding words, etc. Some of these data can be obtained by querying a search engine directly; although this has its drawbacks (ambiguity caused by its non-linguistically-tagged nature, you cannot query for the POS, the sort order is anything but linguistically guided, redundancy...), it also has its advantages (the corpus is huge, constantly updated...). Thus, some services that ease the use of the web as a direct source of linguistic evidence, namely WebCorp (Renouf et al., 2007) or KWiCFinder (Flectcher, 2006), have appeared. They query the APIs of search engines for the words the user enters, download the pages they return and show occurrences of the word in a KWiC way.

Such a service is very interesting for Basque or for any language not rich in corpora, but since they rely on APIs of search engines, they pose the problems we have already stated. So we have built a service called CorpEus (Leturia et al., 2007), which solves these by means of morphological query expansion and language-filtering words. It is available for querying at http://www.corpeus.org.

3.2 Terminology

The Elhuyar Foundation has developed several tools to automatically extract monolingual or multilingual terminology out of different kinds of corpora, using a combination of linguistic and statistical methods.

Erauzterm (Gurrutxaga et al., 2004) is a tool for automatic term extraction from Basque corpora, implemented by the Elhuyar Foundation in collaboration with the IXA group. It has reported F measure results of 0.4229 for multi-word terms and 0.4693 for single word terms, and precision values of up to 0.65 for multi-word terms and up to 0.75 for single word terms for the first 2,000 candidates over a corpus on electricity & electronics.

Elexbi (Alegria et al., 2006) extracts pairs of equivalent terms from Spanish-Basque translation memories. It is based on monolingual candidate extraction in Basque (Erauzterm) and Spanish (Freeling), and consequent statistical alignment and extraction of equivalent pairs. It has reported results of up to 0.9 precision for the first 4,000 candidates processing a parallel corpus of 10,900 segments.

AzerHitz (Saralegi, et al., 2008a; Saralegi, et al., 2008b) is a tool to automatically extract pairs of equivalent terms from Basque-English or Basque-Spanish domain-comparable corpora based on context similarity, obtaining a precision of 58% in top 1 and 79% in top 20 for high-frequency words.

The combination of these terminology extraction tools with the corpora collection tools we have mentioned above, provides some semi-automatic ways of building dictionaries out of the web:

- AutoCorpEx collects Basque specialized corpora from the web, and then we obtain lists of terms in Basque by applying Erauzterm to them.
- Co3 can gather English-Basque comparable corpora out of the web, and by applying AzerHitz to them we obtain English-Basque terminology lists.

- PaCo2 will, in a near future, collect Spanish-Basque parallel corpora from the web and then Elexbi will extract Spanish-Basque terminology from them.

The next section describes some experiments we have conducted using the first two, since the corpus collection tool of the third approach is still under development.

3.3 Ontologies

There is also an ongoing project for automatically extracting specialized terminology out of a Basque corpus, in order to automatically (or semi-automatically) enrich existing concept taxonomies such as WordNet, or in order to build domain-specific ontologies. The specialized corpora to be used in this project can also be collected automatically out of the web.

4. Experiments

In this section we will show some experiments we have performed to use the web as "raw material" to build language resources such as corpora and term lists. Our first task will be to explore the possibilities that the web offers for the compilation of terminological dictionaries in Basque, via automatic term extraction from web-corpora. We will use AutoCorpEx for collecting specialized web corpora in Basque and Erauzterm as the Basque term extraction tool. In the second experiment, we enter the field of comparable corpora, and present some experiments that envisage the construction of multilingual terminological resources for language pairs with scarce parallel corpora such as Basque. We use Co3 for compiling the domain-comparable corpora and AzerHitz for extracting bilingual terminology out of them. The experiment aims to improve the performance of the terminology extraction by using the web for collecting additional data on the fly to improve context-similarity computation.

4.1 Monolingual specialized web corpora

The goal of the first experiment is to evaluate the domain precision of the web corpora built with Co3 and of the term lists extracted out of them with Erauzterm.

4.1.1. Design

We collected three specialized corpora in the domains of Computer Science, Biotechnology and Atomic & Particle Physics. The collection of the corpora from the Internet did not have a target size, because the Internet in Basque is not as big as that in other languages, and the number we would want to collect for a particular domain might not exist. So we simply launched the collecting processes and stopped them when the growing speed of the corpora fell to almost zero, thus obtaining corpora that were as large as possible.

Then we applied the terminology extraction process to the corpora and obtained the three term lists. These lists were automatically validated against a recently compiled specialized dictionary, ZT Hiztegia or Basic Dictionary of Science and Technology (http://zthiztegia.elhuyar.org), which contains 25,000 terms, and the online version of Euskalterm, the Basque Public Term Bank (http://www1.euskadi.net/euskalterm/indice_i.htm). The terms not found in those terminological databases were manually validated by experts up to a certain number.

Table 1 shows the size of the corpora obtained, the number of terms extracted and the number of terms validated manually or by the dictionary, for each of the three domains.

4.1.2. Evaluation and results

Firstly, we evaluated the domain precision of the lists obtained from the Internet, by analyzing the distribution of the terms across the domains, taking the domains of the specialized dictionary as a reference. The results of this evaluation are shown in Figure 1.

We can observe that all three lists show peaks in or around their respective domains, which proves that the corpora are indeed specialized to some extent and that the term lists automatically extracted belong mainly to the desired domains.

On the other hand, the Biotechnology corpus appears to be the less specialized one, as its distribution is flatter than the others'. Besides, in that corpus and especially in the Computer Science one, the presence of terms not belonging to the area of science and technology is remarkable. The explanation for this could be that they both are technology domains, and hence are closely related to their application areas; not surprisingly, terms from those applications areas occur in those texts more frequently than in pure science documents.

Figure 2 shows the domain precision of the term extraction for each corpus (relative to valid terms). A distinction between General Physics and Atom & Particle Physics has been made. An explanation for the fact that precision results are considerably better for the former could be that many general terms in Physics occurred along with atomic and particle terminology. We may be able to understand this if we take into account the fact that most of the texts are not the product of communication among specialists, but of popular science or teaching materials.

Regarding recall relative to the ZT Hiztegia (Figure 3), the best results are obtained for Atomic & Particle Physics, while the recall for Biotechnology is the lowest. The overall conclusion could be that the three web corpora are lacking representativeness, and are not good enough for compiling a quality dictionary. There is no single possible explanation for that. For example, in the case of Atomic & Particle Physics, out of the 474 terms included in the dictionary, 150 were not extracted from the web corpus (31.64%). We checked the presence of those 150 terms in the Internet, and 42 of them were not retrieved by Google (using CorpEus). 4 terms are in the Internet, but not in the web corpus, and finally, 104 terms in the web corpus were not extracted by Erauzterm (101 occurring only once).

So the main problem is the recall of the Basque Internet itself (Erauzterm could hardly be blamed for not being able to extract 101 terms with f = 1).

One possible explanation for this fact could lie in the current situation of Basque terminology and text production. Although Basque began to be used in Science and Technology thirty years ago, it cannot be denied that there is a given amount of highly specialized terminology that is published *ex novo* in dictionaries, with little document support if any. That could be the reason why several terms chosen by experts and published in the dictionary do not occur or occurred only once in the Internet.

Finally, as we can see in Table 2, the manual validation process provided new terms not included in the dictionary. This suggests that the process proposed could be interesting for enriching or updating already existing specialized dictionaries.

More details and results of this experiment can be found in a paper entirely dedicated to it (Gurrutxaga et al., 2009).

Corpus	**Atomic and Particle Physics**	**Computer Science**	**Biotechnology**
Sample corpus size	32 docs, 26,164 words	33 docs, 34,266 words	55 docs, 41,496 words
Obtained corpus size	320,212	2,514,290	578,866
Extracted term list size	46,972	163,698	34,910
Dictionary validated	6,432	8,137	6,524
First 10,000 candidates	2,827	2,755	2,403
Manually evaluated	869	904	628
Terms	628	512	432
Not terms	241	392	196

Table 1. Corpus and term list sizes obtained for each of the three domains

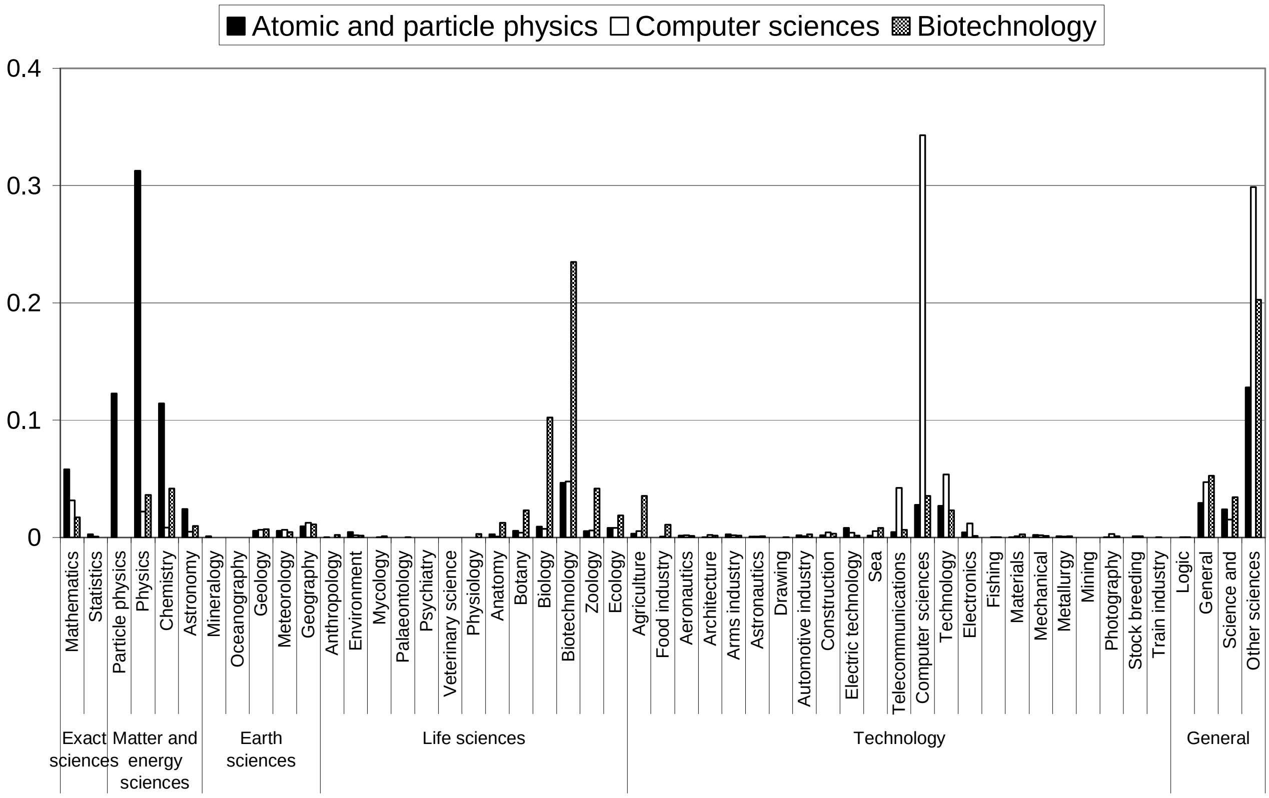

Figure 1. Domain distribution of the extracted term lists

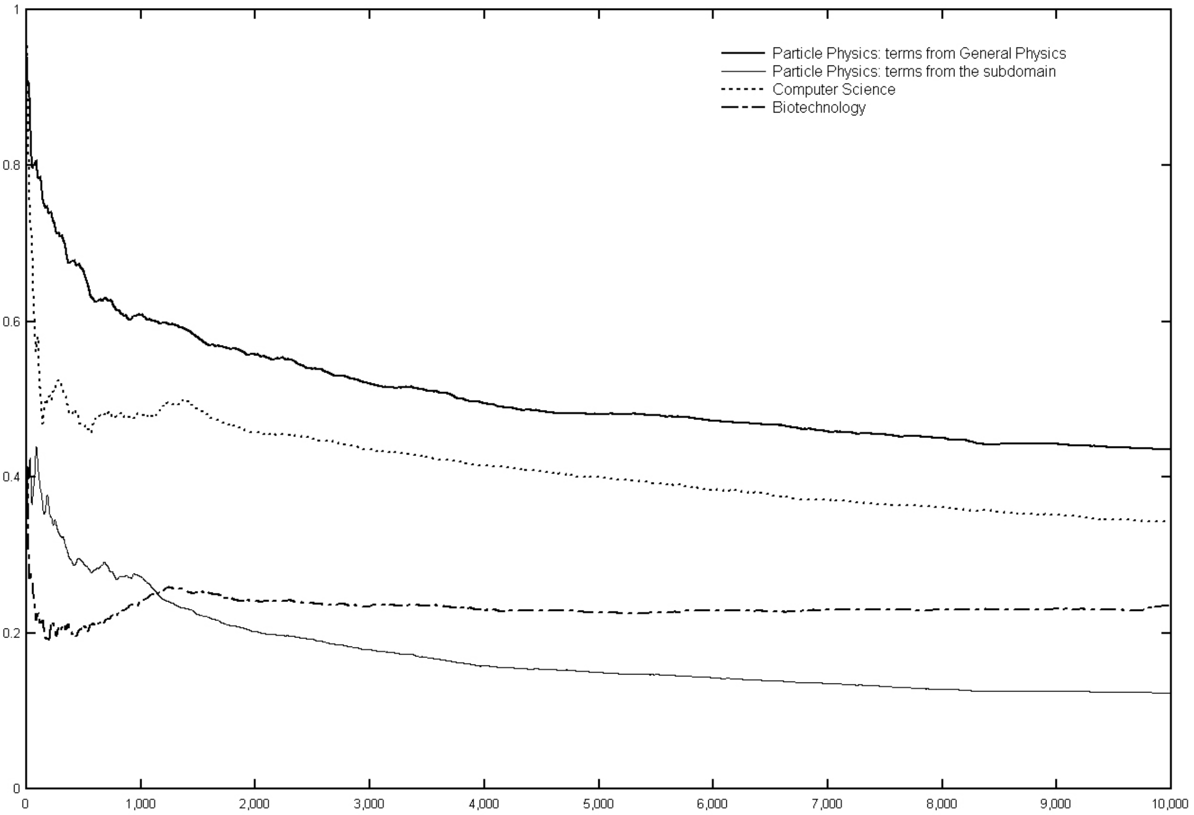

Figure 2. Domain precision of term extraction from each web corpus (relative to validated terms)

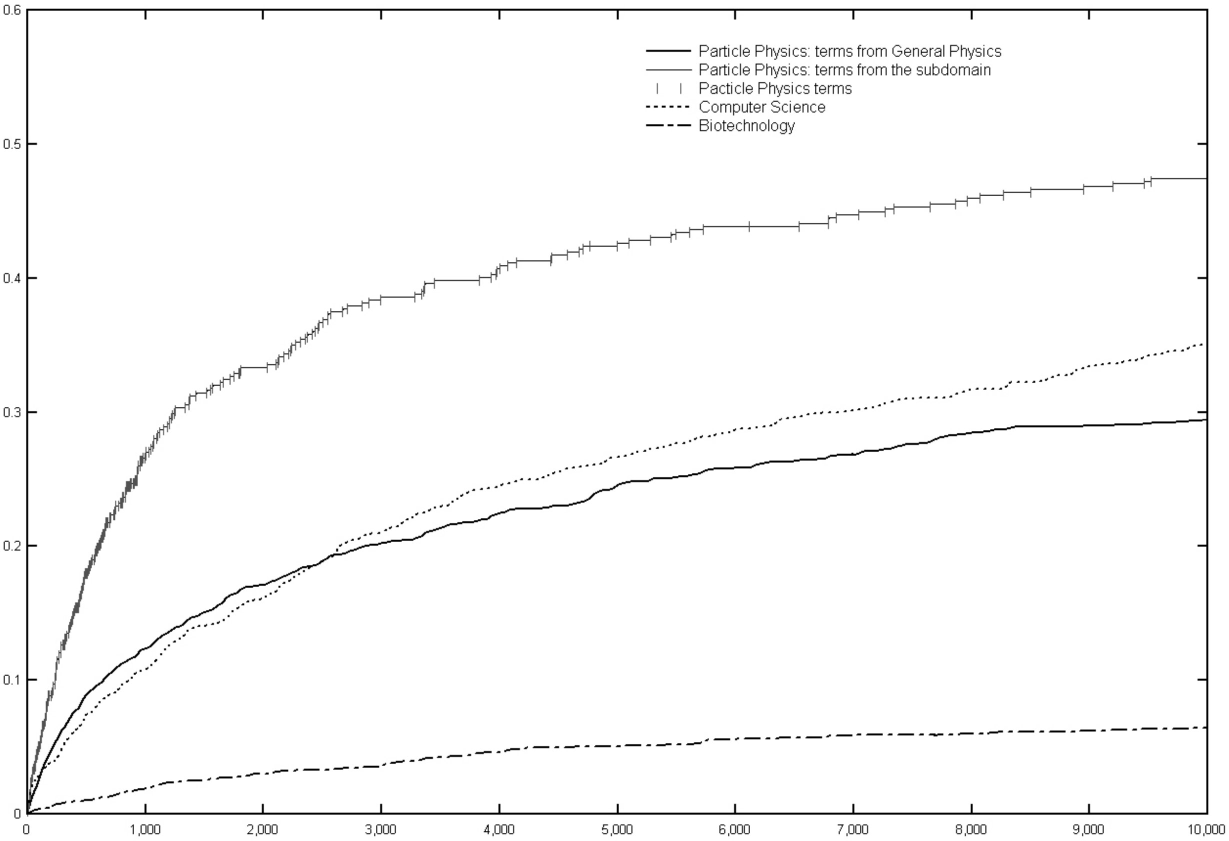

Figure 3. Domain recall of each term extraction

Atomic and Particle Physics		**Computer Science**		**Biotechnology**	
Physics	377	Computer Science	348	Biotechnology	146
Atomic and Particle Physics	109	General	112	Biology	99
Chemistry	56	Telecommunications	22	General	92
Others	86	Others	30	Others	95
Total	628	Total	512	Total	432

Table 2. Distribution of the new terms obtained by manual validation of the candidates extracted from the web corpora

4.2 Multilingual domain-comparable web corpora

This second experiment evaluates the improvement obtained in AzerHitz by enhancing the contexts of words with Internet searches. For this purpose, we have extracted bilingual terminology lists automatically with the AzerHitz tool from a Basque-English comparable corpus in the Computer Science domain automatically collected by Co3. Previous research done within AzerHitz is explained in (Saralegi et al., 2008a; Saralegi et al., 2008b). It must be noted that this research is currently ongoing and that the results presented here are preliminary.

4.2.1. Design

There are several reasons for choosing the Computer Science domain. On the one hand, terminology in this domain is constantly increasing. On the other, it is easy to obtain Computer Science documents from the Internet. Hence, terminology extraction from comparable corpora in this domain offers us a versatility that parallel corpora do not offer, because terminologically updated corpora can be easily obtained from the Internet.

For building the corpus, we provided a sample corpus consisting of 5,000 words for each language and launched the Co3 tool with them. Table 3 shows the size of the subcorpora collected.

In order to automatically extract terminology from comparable corpora, the AzerHitz system is based on cross-lingual context similarity. The underlying idea is that the same concept tends to appear with the same context words in both languages, in other words, it maintains many collocates. The algorithm used by AzerHitz is explained next.

AzerHitz starts the process by selecting those words which are meaningful (nouns, adjectives and verbs), henceforth content words. Each of them is then represented by a “context document”. The context document of a word is composed by the content words appearing in the contexts of the word throughout the whole corpus. Those contexts are limited by a maximum distance to the word and by the punctuation marks. Context documents of all of the target language words are indexed by Lemur IR toolkit as a collection using the Indri retrieval model. To be able to compute the similarity between context documents of different languages, the documents in the source language are translated using a bilingual machine readable dictionary. We try to minimise the number of out-of-vocabulary words by using cognate detection, and ambiguity is tackled by using a first translation approach. To find the translation of a source word, its translated context document is sent as a query to the IR engine which returns a ranking of the most similar documents. In addition, a cognate detection step can be performed over the first ranked candidates. If a cognate is detected, the corresponding candidate will be promoted to the first position in the ranking. This can be useful in some domains in which the presence of loanwords is high.

The main problem of the context similarity paradigm is that the majority of the words do not have enough context information to be represented properly. To mitigate this problem, we propose that the Internet be used as a big comparable corpus. In this way, we expand the contexts of a word obtained from the initial corpus with new context words retrieved from web concordancers such as WebCorp (Renouf et al., 2007) or CorpEus (Leturia et al., 2007) to get a richer representation of the context. The contexts of both source and target language words are expanded. However, expanding all the contexts in the target language is computationally too expensive, and that is why, we only apply the expansion to the first translation candidates ranked by the IR engine.

The expansion may seem as a trivial task, but it has to address certain difficulties. We can not just expand with any context we get, because we may add noisy data. The contexts added must refer to the same sense of the word represented by the corpus contexts. In order to guarantee information with a good quality we use domain control techniques when retrieving contexts from the web concordancers.

4.2.2. Evaluation and results

We have evaluated the increase in performance obtained in AzerHitz by applying the enhancement of contexts using the web.

The evaluation of the system has been done over a set of 100 words, taken randomly from the corpus and which are not in the dictionary used. The words are translated manually in order to set up the reference for performing an automatic evaluation.

The following setups have been evaluated:

- Baseline: Only contexts obtained from the corpus.
- Baseline + Cognates: Cognate detection is performed on the first 20 ranked candidates.
- WaC: Web contexts expansion is performed.
- WaC + Cognates: Both context expansion and cognate detection among the first 20 ranking candidates are performed (in that order).

Table 4 shows the results of the experiments. Although these are only preliminary results, we can see that the expansion of the contexts using web data outperforms the results achieved when the context alone is retrieved from the corpus. These results show that the expansion helps to represent the word contexts better and, in turn, a better representation helps to compute more accurate context similarity and find correct translations.

We can also observe that adding the identification of cognates among the first 20 ranked candidates greatly improves the precision of the final ranking. The high presence of these kinds of translations accounts for this improvement.

Subcorpus	Words	Documents
Basque	2.6 M	2 K
English	2.6 M	1 K

Table 3. Computer science comparable corpus

Setup	top1	top5	top10	top15	top20
Baseline	0.32	0.54	0.60	0.62	0.66
WaC	0.36	0.56	0.68	0.72	0.72
Baseline + cognates	0.54	0.62	0.62	0.64	0.66
WaC + cognates	0.58	0.66	0.70	0.72	0.72

Table 4. Precision for top rankings

5. Conclusions

A common problem of less resourced languages is that the economic resources devoted to the development of NLP tools are also scarce. So the use of the Internet for building language resources such as corpora and, through them, other resources and NLP tools, is very attractive indeed. Nevertheless, the hypothesis that the Internet is a valuable and profitable source for developing language resources for less resourced languages must be tested in order to set up initiatives and projects with that objective.

It goes without saying that any attempt to build web corpora in a given language is conditioned by the size of the web in the target domains or genres. We consider that the results of the experiments that we have presented for Basque are encouraging. The size of the specialized web corpora we have compiled with our tools and the domain-precision achieved gives us some evidence that the Basque Internet, although not in any way comparable with the webs of major languages, can be large enough in specialized domains to be considered as a data source. Also, the fact that the use of web-derived contexts improves the results of terminology extraction from comparable corpora is further proof of this. This optimism should not hide the fact that, for the time being, some domains and genres may not have enough representation in the web.

In view of all this, the Elhuyar Foundation will go on working with the web as a source of corpora of many kinds and other types of language resources for Basque.

6. References

Alegria, I., Gurrutxaga, A., Saralegi, X., Ugartetxea, S. (2006). Elexbi, a basic tool for bilingual term extraction from Spanish-Basque parallel corpora. In *Proceedings of Euralex 2006*. Torino: Euralex, pp. 159--165.

Baroni, M., Bernardini, S. (2004). BootCaT: Bootstrapping corpora and terms from the web. In *Proceedings of LREC 2004.* Lisbon, Portugal: ELDA, pp. 1313--1316.

Baroni, M., Kilgarriff, A. (2006). Large linguistically-processed Web corpora for multiple languages. In *Proceedings of EACL 2006*. Trent, Italy: EACL, pp. 8--90.

Baroni, M., Ueyama, M. (2006). Building general- and special purpose corpora by Web crawling. In *Proceedings of the 13th NIJL International Symposium*. Tokyo, Japan: NIJL, pp. 31--40.

Baroni, M., Bernardini, S., Ferraresi, A., Zanchetta E. (2009). The WaCky Wide Web: A Collection of Very Large Linguistically Processed Web-Crawled Corpora. *Language Resources and Evaluation Journal*, 43(3), pp. 209--226.

Broder, A.Z. (2000). Identifying and filtering near-duplicate documents. In *Proceedings of Combinatorial Pattern Matching: 11th Annual Symposium.* Montreal, Canada: Springer, pp. 1--10.

Broder, A.Z. (1997). On the resemblance and containment of documents. In *Proceedings of Compression and Complexity of Sequences 1997.* Los Alamitos, CA: IEEE Computer Society, pp. 21--29.

Fletcher, W.H. (2004). Making the web more useful as a source for linguistic corpora. In U. Connor & T. Upton (Eds.), *Corpus Linguistics in North America 2002*. Amsterdam, The Netherlands: Rodopi.

Fletcher, W. H. (2006). Concordancing the Web: Promise and Problems, Tools and Techniques. In M. Hundt, N. Nesselhauf & C. Biewer (Eds.), *Corpus Linguistics and the Web*. Amsterdam, The Netherlands: Rodopi, pp. 25--46.

Gurrutxaga, A., Saralegi, X., Ugartetxea, S., Lizaso, P., Alegria, I., Urizar, R. (2004). A XML-based term extraction tool for Basque. In *Proc. of fourth* *international conference on Language Resources and Evaluation (LREC)*. Lisbon, Portugal: ELRA, pp. 1733--1736.

Gurrutxaga, A., Leturia, I., Pociello, E., Saralegi, X., San Vicente, I. (2009). Evaluation of an automatic process for specialized web corpora collection and term extraction for Basque. In *Proceedings of eLexicography in the 21st century*. Louvain-la-Neuve, Belgium: EURALEX & SIGLEX.

Leturia, I., Gurrutxaga, A., Alegria, I., Ezeiza, A. (2007). CorpEus, a 'web as corpus' tool designed for the agglutinative nature of Basque. In *Proceedings of Web as Corpus 3 workshop*. Louvain-la-Neuve, Belgium: ACL-SIGWAC, pp. 69--81.

Leturia, I., San Vicente, I., Saralegi, X., Lopez de Lacalle, M. (2008a). Collecting Basque specialized corpora from the web: language-specific performance tweaks and improving topic precision. In *Proceedings of the 4th Web as Corpus Workshop*. Marrakech, Morocco: ACL SIGWAC, pp. 40--46.

Leturia, I., Gurrutxaga, A., Areta, N., Pociello, E. (2008b). Analysis and performance of morphological query expansion and language-filtering words on Basque web searching. In *Proceedings of LREC 2008*. Marrakech, Morocco: ELRA.

Leturia, I., San Vicente, I., Saralegi. X. (2009). Search engine based approaches for collecting domain-specific Basque-English comparable corpora from the Internet. In *Proceedings of 5th International Web as Corpus Workshop (WAC5)*. Donostia, Spain: ACL-SIGWAC, pp. 53--61.

Renouf, A., Kehoe, A., Banerjee, J. (2007). WebCorp: an Integrated System for WebText Search. In M. Hundt, N. Nesselhauf & C. Biewer (Eds.), *Corpus Linguistics and the Web*. Amsterdam, The Netherlands: Rodopi, pp. 47--67.

Resnik, P. (1998). Parallel strands: A preliminary investigation into mining the web for bilingual text. In *Proceedings of the Third Conference of the Association for Machine Translation in the Americas on Machine Translation and the Information Soup*. Langhorne, USA: AMTA, pp.72--82,

Saralegi, X., Alegria, I. (2007). Similitud entre documentos multilingües de carácter científico-técnico en un entorno web. *Procesamiento del Lenguaje Natural*, 39, pp. 71--78.

Saralegi, X., Leturia, I. (2007). Kimatu, a tool for cleaning non-content text parts from HTML docs. In *Proceedings of the 3rd Web as Corpus workshop*. Louvain-la-Neuve, Belgium: Presses universitaires de Louvain, pp. 163--167.

Saralegi, X., San Vicente, I., Gurrutxaga, A. (2008a). Automatic Extraction of Bilingual Terms from Comparable Corpora in a Popular Science Domain. In *Proceedings of Building and using Comparable Corpora workshop*. Marrakech, Morocco: ELRA.

Saralegi, X., San Vicente, I., López de Lacalle, M. (2008b). Mining Term Translations from Domain Restricted Comparable Corpora. *Procesamiento del Lenguaje Natural*, 41, pp. 273--280.

Sebastiani, F. (2002). Machine learning in automated text categorization. *ACM Computing Surveys*, 34(1), pp. 1--47.

Sharoff, S. (2006). Creating General-Purpose Corpora Using Automated Search Engine Queries. In Marco Baroni & Silvia Bernardini (Eds.), *WaCky! Working Papers on the Web as Corpus*. Bologna, Italy: Gedit Edizioni, pp. 63--98.